Tartu Ülikool

Loodus- ja täppisteaduste valdkond

Matemaatika ja statistika instituut

Indrek Polding

Krediidiriski hindamine logistilise regressiooni mudeli abil

Matemaatilise statistika eriala

Bakalaurusetöö (9 EAP)

Juhendaja: prof. Kalev Pärna

Tartu 2018

Krediidiriski hindamine logistilise regressiooni mudeli abil

Bakalaureusetöö

Indrek Polding

Lühikookuvõte. Finantsettevõtete üheks suureks tegevusvaldkonnaks on laenude väljastamine. Selles protsessis on vaja hinnata laenutaotleja krediidiriski, et kindlaks määrata, kui suur on tõenäosus, et laenutaotleja tulevikus oma laenu korralikult tagasi maksaks. Prognoosi tulemusena klassifitseeritakse laenutaotlejad kahte gruppi: head ja halvad kliendid. Kuna uuritav tunnus on binaarne, siis üheks enam kasutatavaks meetodiks on logistiline regressioon. Töö teoreetilises osas antakse ülevaade krediidiriskist ja tutvustatakse töös kasutatud metoodikat. Praktilises osas antakse ülevaade andmestikus olevates tunnustest ja luuakse mudel nii tasakaalustatud kui ka tasakaalustamata andmete põhjal ja interpreteeritakse parimat saadud mudelit.

Märksõnad: krediidirisk, logistiline regressioon, üldistatud lineaarsed mudelid

CERCS teaduseriala: P160 Statistika, operatsioonanalüüs, programmeerimine, finants- ja kindlustusmatemaatika

Estimating a credit risk by logistic regression

Bachelor’s thesis

Indrek Polding

Abstract. One of the big business areas for financial corporations is the issuance of loans. In this process, it is necessary to evaluate the credit applicant's credit risk in order to determine how likely a borrower is to repay its loan properly in the future. As a result of the forecast, loan applicants are classified into two groups: good and bad customers. Since the investigated character is binary, one of the more usable methods is logistic regression. The theoretical part of the work gives an overview of credit risk and introduces the methodology used in the work. In the practical part, an overview of the characteristics of the data is given and the model is created on the basis of both balanced and unbalanced data and the best model is interpreted.

Keywords: credit risk, logistic regression, generalized linear models

CERCS research specialisation: P160 Statistics, operation research, programming, actuarial mathematics

Sisukord

Sissejuhatus

Finantsettevõtete üheks suureks tegevusvaldkonnaks on laenude väljastamine. Kuna ettevõtte eesmärgiks on kasumi maksimeerimine, siis on vajalik laenutaotlejate klassifitseerimine headeks ja halbadeks klientideks. Sellega saadakse vältida suuri kahjusid halbadele klientidele mitte laenu andmisega kui ka teenida kasumit määrates laenusaajatele erinevaid riskimäärasid. Klassifitseerimine toimub krediidiriski hindamise abil. Krediidiriski hindamisel prognoositakse kui suure tõenäosusega võib laenusaaja laenu tagasimaksmisel makseraskustesse sattuda. Tõenäosuste arvutamine toimub laenusaajate andmete põhjal, kus lisaks kliendi isiklikule ja laenuga seotud infole on iga kliendi kohta lisatud, kas tegemist on halva või hea kliendiga.

Töö eesmärgiks on luua mudel, mille abil saaks prognoosida tõenäosust, et klient osutub heaks kliendiks. Mudeli loomisel tahetakse ka välja selgitada, millised tunnused seda tõenäosust mõjutavad. Samuti on töö eesmärgiks uurida andmete tasakaalustamise probleemi ehk kuivõrd oluline on heade ja halbade klientide võrdne esindatus valimis kui kasutatakse logistilist regressiooni Sellele küsimusele on kirjanduses sageli viidatud. [7]

Töö esimeses osas antakse ülevaade krediidiriskist, teises osas tutvutakse töös kasutatud metoodikaga: antakse ülevaade logistilise regressiooni mudelist ja selle headuse näitajatest, kirjeldatakse tunnuste esmaanalüüsis kasutatavat hii-ruut testi ning kirjeldatakse andmete tasakaalustatuse probleemi. Töö kolmandas osas teostatakse andmestikus olevatele tunnuste esmaanalüüs ja uuritakse, kuidas käsitletav argumenttunnus ja funktsioontunnus omavahel seotud on. Samuti kirjeldatakse logistilise regressiooni mudeli hindamise käiku, võrreldakse tasakaalustamata ja tasakaalustatud andmete põhjal loodud regressioonmudeleid ja interpreteeritakse lõplikku mudelit.

Bakalaureusetöö analüütiline osa on teostatud rakendustarkvaraga R.

Töö koostaja soovib tänada juhendajat professor Kalev Pärnat heade ja edasiviivate soovituste eest.

1 Ülevaade krediidiriskist

Krediidiskooringut kasutatakse selleks, et hinnata, kui tõenäoliselt kliendid ei täida oma kohustust laenu tagasi maksta. Tõenäosust, et laenusaaja jääb võlgu, hinnatakse laenu andmise otsustusprotsessi ajal ja selleks kasutatakse infot, mida laenutaotleja annab. Tõenäosuse arvutamise tulemusena tehakse otsus, kas laenu anda või mitte. Täpsest laenutaotlejate klassifitseerimisest headeks ja halbadeks klientideks saavad kasu nii laenuandjad (kasumi suurendamine ja kahjude vähendamine kui ka laenutaotleja (finantsilise ülekoormuse vältimine). Traditsiooniliselt põhines otsuse tegemine laenuandja intuitsioonil, mis toetus laenutaotleja varasemal käitumisel. Majandusliku surve ja uute tehniliste lahenduste tulemusena on üle mindud statistiliste mudelite kasutamisele, mida nimetatakse krediidiskooringu süsteemideks. Selle süsteemi järgi on klient madala riskiga, kui ta omab suurt tõenäosust osutuda heaks kliendiks ja kõrge riskiga, kui ta omab madalat tõenäosust osutuda heaks kliendiks. [1]

Finantskompaniid püüavad üha tihedamini säilitada oma kliente pakkudes neile uusi tooteid ja tõhusamaid teenuseid. Riskihalduse osakonnal on vaja selgitada välja madalama riskiga kliendid, kellele neid võimalusi pakkuda. Selleks on vaja arendada välja strateegia, et kindlaks teha, kellele paremat kohtlemist pakkuda. Samuti on vaja kindlaks teha kõrgema riskiga kliendid ja leida nendega tegelemiseks kõige efektiivsem viis, et vähendada kulusid. Riskiskooringu kaardid pakuvad efektiivset ja empiiriliselt tuletatud lahendust ärivajaduste rahuldamiseks. Skoorimismetoodika võimaldab objektiivset võimalust riski hindamiseks ja järjepidevat lähenemist tagamaks, et süsteemi vead oleks võimalikult väikesed. [2]

Varasemalt ostsid finantsettevõtted krediidiskooringu süsteemi sisse mõnelt teenusepakkujalt. Tehnoloogia arenedes hakkas levima praktika, et süsteemid loodi majasiseselt, et vähendada kulusid ja muuta krediidiskooringu süsteemide loomine paindlikumaks ja kiiremaks. Rakendustarkvarade kättesaadavamaks muutumine ja lihtsamad andmete ladustamise võimalused on muutnud skooringusüsteemi loomise odavamaks, kiiremaks ja paindlikumaks, kuna ei olnud enam vaja nii välja arenenud programmeerimisoskust ja ettevõtted said ära kasutada oma paremaid teadmisi andmete sisu ja ettevõtlusalase teabe kohta. [2]

Krediidiskooringu süsteemide põhjal tehtud analüüside toel saab välja töötada uusi strateegiaid, mis aitavad maksimeerida kasumit ja minimiseerida suuri võlgasid. Mõned riskantsete klientidega tegelemise strateegiad on:

- madalama krediidilimiidi määramine,
- laenu intressimäära tõstmine,
- lisada klient jälgimisnimekirja,
- paluda laenutaotlejal anda tagatist kommunaalteenuste eest,
- loobuda laenu andmisest.

Madala riskiga klientide puhul pakutakse võimalust suurendada krediiti krediitkaardil, pakutakse lisatooteid, võimaldatakse paremaid intressimäärasid, lubatakse minna üle lubatud krediidimäära krediitkaardil. [2]

Hinnangud on näidanud, et logistiline regressioon on olnud väga edukas meetod efektiivse skooringumudeli loomiseks. Logistiline regressioon suudab hõlmata turumajanduse erinevaid tunnuseid ja kindlaks teha kõige suurema mõjuga tunnused, et pank saaks kindlaks teha kliendid, kes kõige suurema tõenäosusega võivad sattuda makseraskustesse. [3]

Logistiline regressioon on ka selle tõttu laialt kasutusel, et uuritav tunnus on binaarne — ennustatakse seda, kas klient on hea või halb. Samuti kasutatakse krediidiriskide hindamisel lineaarset regressiooni, diskriminantanalüüsi, CART mudeleid ja teisi masinõppe meetodeid. Kuigi loetletud meetodid on laialt kasutusel, siis on nende kohta siiski vähe teoreetilist infot, sest kehtib konfidentsiaalsusnõue. Loodud krediidiskooringu süsteemi ja välja töötatud meetodite avaldamine võib tuua ettevõttele finantsilist kahju. [1]

2 Kasutatav metoodika

2.1 Logistilise regressiooni mudel

Antud alapeatükk põhineb loengukonspektil [4], kui pole viidatud teisiti.

Tihti pakub analüütikule huvi diskreetne uuritav tunnus, millel on ainult kaks võimalikku väärtust. Uuritava tunnuse puhul on tegemist binaarse tunnusega Y, mille kodeerimisel kasutatakse harilikult väärtusi 1 ja 0, kus 1 tähistab sündmuse esinemist.

Logistilise regressioonmudeli korral on uuritav tunnus Y Bernoulli jaotusest $Y \sim Be(1, p)$, kus p on meid huvitava sündmuse tõenäosus. Antud töös on uuritavaks tunnuseks see, kas tegemist on hea või halva kliendiga, mudeli abil ennustatakse seda, kas tegemist on hea kliendiga (Y=1).

Binaarse tunnuse korral on kasutusel $logit$-seosefunktsioon, kus $logit(p) = ln\frac{p}{1-p}$, kus $\frac{p}{1-p}$ on sündmuse esinemise šanss (sündmuse esinemise tõenäosuse ja sündmuse mitteesinemise tõenäosuse suhe).

Logistilise mudeliga hinnatakse seega šansi logaritmi:

$$ln\frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k,$$

kus $p = P(Y = 1)$ on sündmuse esinemise tõenäosus, $\beta_0, \beta_1, \dots, \beta_k$ on mudeli tundmatud parameetrid ja $x_1, x_2, \dots, x_k$ seletavad tunnused ehk argumenttunnused.

Logistilises regressioonis kasutatakse tundmatute parameetrite β_i hindamiseks suurima tõepära meetodit. Kasutatakse valimi tõepärafunktsiooni, mille kuju sõltub uuritava tunnuse jaotusest ja mida tähistatakse $L(Y, p)$, kus Y on valim ja p otsitav parameeter.

$$L(Y, p) = \prod_{i=1}^{n} P(Y_i; p_i) = \prod_{i=1}^{n} [p_i^{Y_i}(1 - p_i)^{1-Y_i}]$$

Suurima tõepära hinnangu korral leitakse selline parameetri väärtus, mille korral tõepärafunktsioon saavutab maksimumi. Tehnilise töö lihtsustamiseks on kasutusele võetud tõepärafunktsiooni logaritmitud kuju ehk logaritmiline tõepärafunktsioon $l = \ln(Y, p)$, mis saavutab maksimumi samas kohas. [9]

Olgu $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ mudeli kordajate hinnangud, mis on leitud suurima tõepära meetodil. $Logit$ seosest saab avaldada huvipakkuva sündmuse esinemise tõenäosuse prognoosi:

$$\hat{p} = \frac{e^{\eta}}{1+e^{\eta}},$$

kus

$$\eta = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k.$$

Iga uuritava tunnuse kohta saab sõnastada hüpoteesipaari selle kohta, kas argumenttunnus i ($i \in \{1, \dots, k\}$) mõjutab laenusaaja tõenäosust olla hea klient:

$$H_0: \beta_i = 0,$$

$$H_1: \beta_i \neq 0.$$

Peatume lühidalt parameetrite tõlgendamisel. Kui hinnatud parameeter $\hat{\beta}_i$ on positiivne, siis on argumenttunnuse ja uuritava tunnuse vahel samasuunaline seos (argumendi väärtuse suurenemise korral suureneb ka laenusaaja tõenäosus olla hea klient). Kui aga hinnatava parameetri $\hat{\beta}_i$ märk on negatiivne, siis on argumenttunnuse ja uuritava tunnuse vahel vastassuunaline seos (argumendi väärtuse suurenemise korral väheneb laenusaaja tõenäosus olla hea klient). Üksiku parameetri $\hat{\beta}_i$ väärtust interpreteeritakse järgmiselt: argumendi x_i suurenemisel ühe ühiku võrra suureneb šansside suhe $e^{\widehat{\beta}_i}$ korda eeldusel, et teiste argumenttunnuste väärtused ei muutu. Kui argument suureneb c ühiku võrra, siis šansside suhe muutub $e^{c\widehat{\beta}_i}$ korda. Vabaliikme interpretatsioon on võimalik juhul, kui $x = 0$ on argumendi võimalik väärtus. Sel juhul on positiivse vabaliikme korral sündmuse esinemise tõenäosus $p > 0.5$. Negatiivset vabaliiget interpreteerida ei saa.

2.2 Mudeli headuse näitajad

Selles bakalaureusetöös on logistilise mudeli headuse näitajatena kasutatud Akaike informatsioonikriteeriumi, ROC-kõvera alust pindala ja ruutkeskmist viga. Mudeli sobivuse kontrolliks kasutati Hosmer-Lemeshow' testi.

Akaike informatsioonikriteerium (AIC) saadakse hälbimusele teatud parandusliikme lisamisel. [4]

Ruutkeskmine viga arvutatakse logistilise regressioonimudeli korral valemist

$$MSE = \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{n},$$

kus n on valimi maht, y_i uuritava tunnuse väärtused ja $\hat{y}_i$ mudeli abil prognoositud väärtused. [4]

ROC-kõvera ja Hosmer-Lemeshow' testi kirjeldamiseks on kasutatud teost [5].

ROC-kõvera alune pindala (AUC) on üks enim kasutatavaid logistilise regressiooni headuse näitajaid. Mudeli abil leitakse igale vaatlusele sündmuse esinemise tõenäosus. Binaarse sündmuse prognoosimiseks tuleks valida lävend, mille alusel määratakse prognoos, kas ennustada sündmuse toimumist või mittetoimumist. Prognoosimisel määratakse lävendist suurema tõenäosusega prognoosile väärtuseks 1, madalama tõenäosusega prognoosile aga väärtuseks 0.

Klassifitseerimisel saab vaadelda kahte näitajat, mida soovitakse maksimiseerida mudeli koostamisel, nendeks on spetsiifilisus ja tundlikkus. Spetsiifilisuse korral on tegemist õigesti prognoositud negatiivsete vaatluste arvu (halbade klientide arvuga töös loodud mudeli puhul) ja kogu negatiivsete vaatluste arvu jagatis. Tundlikkus on õigesti prognoositud positiivsete vaatluste arvu (heade klientide arvu) ja kogu vaatluste arvu jagatis. Seega näitab (1 – spetsiifilisus) valepositiivsete vaatluste määra ja tundlikkus tõeselt positiivsete vaatluste määra.

ROC-kõvera konstrueerimisel kuvatakse horisontaalteljele valepositiivsete vaatluste määr ja vertikaalteljele tõeselt positiivsete vaatluste määr. Tundlikkus ja spetsiifilisus sõltuvad lävendi valikust.

ROC-kõvera alune pindala iseloomustab mudeli võimet korrektselt prognoosida sündmuse esinemise tõenäosust. Vastavalt ROC-kõvera aluse pindala väärtusele võib logistilised mudelid jagada järgmistesse klassidesse:

- $AUC = 0{,}5$, tegemist on juhusliku mudeliga, korrektne prognoosivõime puudub,
- $0{,}5 < AUC < 0{,}7$, tegemist on halva mudeliga,
- $0{,}7 \leq AUC < 0{,}8$, tegemist on aktsepteeritava mudeliga,
- $0{,}8 \leq AUC < 0{,}9$, tegemist on väga hea mudeliga,
- $AUC \geq 0{,}9$, tegemist on suurepärase mudeliga.

Hosmer-Lemeshow' testi kasutatakse selleks, et uurida, kas loodud mudel sobib andmetega. Vaatlused järjestatakse prognoositud sündmuse esinemise tõenäosuste järgi ja jagatakse kümneks grupiks ning leitakse teststatistik kujul:

$$C = \sum_{k=1}^{10} \frac{(O_k - N_k\overline{\pi_k})^2}{N_k\overline{\pi_k}(1-\overline{\pi_k})},$$

kus O_k on huvipakkuva sündmuste arv grupis k, N_k on vaatluste arv grupis k ja $\overline{\pi_k}$ on prognoositud tõenäosuste keskmine grupis k. Teststatistik on hii-ruut jaotusega vabaduseastmete arvuga kaheksa. Mudeli andmetega sobivuse testimiseks kasutatakse järgmist hüpoteesipaari:

H_0: mudel sobib andmetega,

H_1: mudel ei sobi andmetega.

2.3 Hii-ruut test

Järgnev alapeatükk põhineb raamatul [6].

Hii-ruut test võrdleb realiseerunud sagedusi oodatavatega. Uuritakse kahte tunnust X ja Y kindlaks tegemaks, kas nad on omavahel sõltuvad. Nullhüpotees väidab sõltumatust:

$$H_0: p_{ij} = p_{i.}p_{.j}, i = 1,2, \dots, I, j = 1,2, \dots, J.$$

Testimiseks kasutatakse Pearsoni hii-ruut statistikut. Oodatavad sagedused leitakse nullhüpoteesi eeldusel:

$$m_{ij} = E(n_{ij}) = np_{ij} = np_{i.}p_{.j},$$

kus $n_{ij} \sim B(n, p_{ij})$. Kuna ka marginaalsed tõenäosused pole harilikult teada, siis hinnatakse need valimist:

$$\hat{p}_{i.} = \frac{n_{i.}}{n}, \hat{p}_{.j} = \frac{n_{.j}}{n}.$$

Saame oodatavate sageduste hinnangud:

$$\hat{m}_{ij} = n\hat{p}_{i.}\hat{p}_{.j} = \frac{\hat{n}_{i.}\hat{n}_{.j}}{n}.$$

Hii-ruut statistiku kuju avaldub järgmiselt:

$$H = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}}.$$

Statistiku H asümptootiline jaotus nullhüpoteesi eeldusel on:

$$H \sim \chi^2(\nu),$$

kus vabadusastmete arv kujuneb järgmiselt:

$$\nu = IJ - 1 - (I-1) - (J-1) = (I-1)(J-1).$$

Siin IJ on juhusliku suuruste n_{ij} arv summas H, neil on üks kitsendus (summeeruvus n-ks). Valimist hinnatakse $I-1$ tunnuse X marginaalset tõenäosust $n_{i.}$, üks $n_{i.}$ on määratud teistega, sest summa on 1. Valimist hinnatakse ka $J-1$ teise tunnuse marginaalset tõenäosust. Kui H_0 kehtib, see tähendab tunnused X ja Y on sõltumatud, siis on realiseerunud n_{ij} ja oodatavad $\widehat{m}_{ij}$ lähedased ja statistiku H väärtus väike. Suur H viitab sõltuvusele. Test tehakse olulisustõenäosuse abil või jaotuse täiendkvantiili abil.

2.4 Andmete tasakaalustamise probleem

Järgmine alapeatükk põhineb artiklil [7], kui pole viidatud teisiti

Finantsvaldkonnas kasutatavates andmestikes juhtub tihti, et binaarse uuritava tunnuse puhul on ühtedega tähistatud väärtuste arv palju väiksem kui nullidega tähistatud väärtuste arv või vastupidi ehk andmed on tasakaalust väljas. Uuritava tunnuse võrdne esindatus on pigem erand kui reegel.

Logistilises regressiooni korral kasutatakse nii tasakaalustatud kui ka tasakaalustamata andmestikke. Tavaliselt andmete tasakaalustamisel võetakse mingi reegli järgi (enamasti juhuslikult) võrdne arv vaatlusi nii nullide kui ühtede seast algsest andmestikust ja seejärel luuakse tasakaalustatud andmestiku baasil logistilise regressiooni mudel. Andmete balansseerimine on praktikas tihti kasutusel, kuid alati ei selgitata selle kasutamise põhjuseid. Tasakaalustamise põhjuste kohta on erinevaid teooriaid:

- tasakaalustamata andmestiku põhjal loodud mudeli korral on kallutatud ainult vabaliikme hinnang (Xie ja Manski (1989));
- tasakaalustamata andmestiku põhjal loodud mudel on kallutatud ainult väikeste valimimahtude korral (Schaefer (1983) ning Scott ja Wild (1986));
- tasakaalustamata andmestiku põhjal loodud mudeli korral on kõik kordajate hinnangud kallutatud (Maddala (1992), King ja Zeng (2001));

- tasakaalustatud andmete põhjal loodud mudeli prognoosid on täpsemad (Maggini (2006) ja Komori(2016)).

Viidatud uuringus analüüsiti hinnatud parameetrite statistilisi omadusi (kallutatus ja hajuvus) ning mudeli täpsust. Uuring näitas, et andmete balansseerimine vähendab hinnatud kordajate hajuvust ja kallutatust. Uuringust selgus samuti, et kui võtta andmestikku proportsionaalselt võrdselt nullisid ja ühtesid, siis prognoosi täpsus suureneb nende väärtuste puhul, mille proportsioon algses andmestikus oli väiksem ja väheneb nende puhul, mille proportsioon oli suurem.

Samas võib kirjanduses kohata ka seisukohti, et logistiline regressioon ei ole andmete balansseerituse suhtes tundlik. [8] Antud töös on logistilist regressiooni rakendatud nii lähteandmetel kui ka balansseeritud andmetel.

3 Laenusaajate logistiline regressioonimudel

3.1 Andmestiku kirjeldus

Käesolevas bakalaureusetöös kasutatava andmestiku puhul on tegu laenusaajate andmestikuga. Tegu on fragmendiga reaalsest andmestikust. Andmestikus on peale puuduvaid väärtusi sisaldavate ridade eemaldamist kokku 3998 vaatlust ja 17 tunnust. Andmestikus on järgmised tunnused:

- laenusaaja staatus, kus 1 – hea, 0 – halb,
- laenusaaja sugu, kus M – mees, F – naine,
- laenusaaja vanus aastates,
- laenusaaja elukoha maakonna nimetus,
- laenusaaja emakeel, kus est – eesti keel, rus – vene keel,
- laenusumma eurodes,
- laenuperiood päevades,
- laenusaaja kuine sissetulek eurodes,
- laenusaaja kuine väljaminek eurodes,
- laenusaaja pereseis,
- laenusaaja haridustase,
- laenusaaja töökogemus,
- laenusaaja laste arv,
- laenusaaja omanduses olevate kinnisvaraobjektide arv,
- laenusaaja aktiivsete maksehäirete arv,
- laenusaaja lõpetatud maksehäirete arv,
- laenusaaja maksehäirete arv kokku.

Laenusaajate andmestikus on hea staatusega kliente 2859 (71,5%) ning halva staatusega kliente 1139 (28,5%). Seega pole andmestik ideaalselt tasakaalus, kuid pole ülearu palju ka tasakaalust väljas.

3.2 Tunnuste esmane analüüs ja teisendamine

Sugu

Tabel 1. Staatuse ja soo omavaheline jaotus

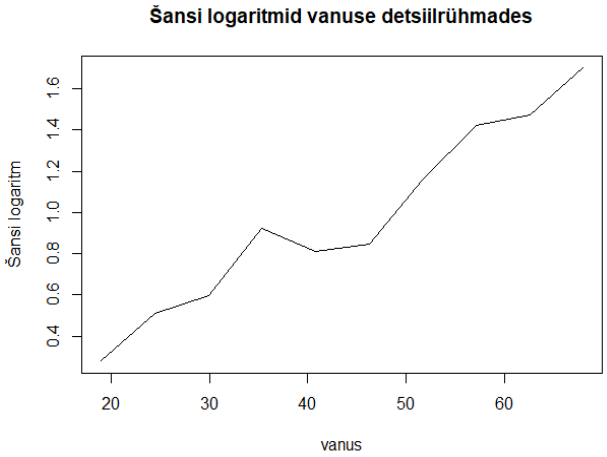
Staatus/sugu	F	M	Kokku
0	506 (25,0%)	633 (32,1%)	1139
1	1521 (75,0%)	1338 (67,9%)	2859
Kokku	2027	1971	3998

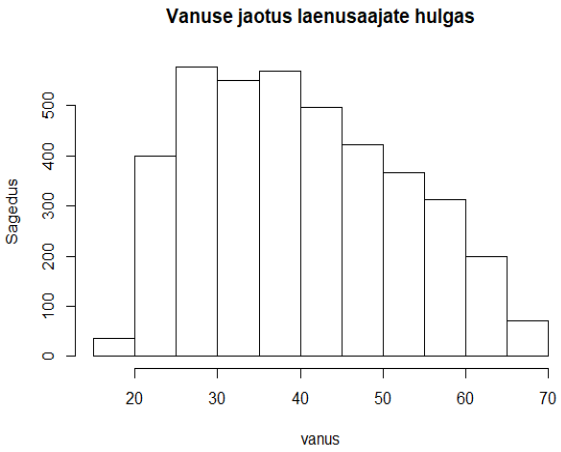
Andmestikku on mehi ja naisi sattunud peaaegu võrdselt (naisi 2027 ja mehi 1971). Tabelist 1 näeme, et naiste seas on häid kliente rohkem (75,0%) kui meeste seas (67,9%). Kontrolliti hii-ruut testiga, kas seos soo ja staatuse vahel on ka statistiliselt oluline. Testi tulemuseks saadi 24.745 (p-väärtus $6{,}44*10^{-7}$). Seega on staatus ja sugu omavahel sõltuvad, mis tähendab, et mehed on suurema tõenäosusega makseraskustesse sattunud.

Vanus

Laenusaajate keskmine vanus on 40,6 aastat, miinimumväärtuseks 19 aastat ja maksimumväärtuseks 68 aastat. Jooniselt 1b võib näha tunnuse vanus jaotust. Kõige rohkem on valimisse sattunud vaatlusi vanuserühmast 25-30, mediaaniks on 39,0 aastat, mis viitab samuti sellele, et keskmisest nooremaid inimesi on valimis rohkem.

Vanuse ja staatuse seose uurimiseks arvutati šansi logaritmid detsiilrühmades, mille jaoks leiti igas vanuserühmas tunnuse staatus keskväärtus ja jagati see staatuselt halbade klientidega osakaaluga vanuserühmades. Joonise 1a põhjal võib arvata, et tunnuste staatus ja vanus vahel esineb ligikaudu lineaarne seos.





Joonis 1a. Šansi logaritmid vanuse detsiilrühmades

Joonis 1b. Vanuse jaotus laenusaajate hulgas

Maakond

Tunnust maakond modifitseeriti selle järgi, kas inimene kuulub Harju maakonda või mitte kuna umbes pooled andmestikus olnud laenusaajatest kuulusid Harju maakonda. Tabelist 2 näeb modifitseeritud tunnuse maakond jaotust:

Tabel 2. Staatuse ja maakonna omavaheline jaotus

Staatus/maakond	Harju	muu	Kokku
0	566 (25,6%)	573 (32,0%)	1139
1	1644 (74,4%)	1215 (68,0%)	2859
Kokku	2210	1788	3988

On näha, et enamik valimisse sattunud laenusaajatest on Harjumaa elanikud (55,4%). Samuti on märgata, et Harju maakonnas on hea staatusega kliente rohkem (74,4%), kui teistes maakondades elavatel inimestel (68,0%). Hii-ruut testi põhjal kontrolliti, kas see seos on statistiliselt oluline: statistiku väärtuseks saadi 19,781 ja p-väärtuseks $8.682 * 10^{-6}$. Seega on maakond ja staatus sõltuvad ehk võib öelda, et Harju maakonna elanikud maksavad paremini laene tagasi.

Keel

Andmestikus on kahte emakeelt kõnelevaid laenusaajaid. Nendeks keelteks on eesti ja vene keel. Tunnuste staatus ja keel jaotus on ära toodud järgnevas tabelis:

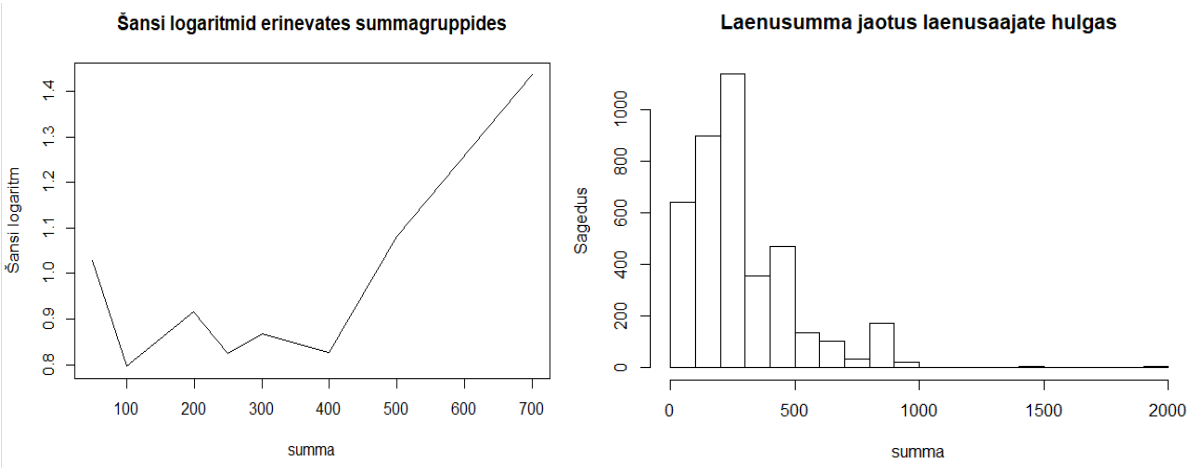
Tabel 3. Staatuse ja emakeele omavaheline jaotus

Staatus/keel	eesti	vene	Kokku
0	642 (29,7%)	497 (27,0%)	1139
1	1518 (70,3%)	1341 (73,0%)	2859
Kokku	2160	1838	3998

Tabelist 3 näeme, et enamik valimisse sattunud inimestest räägivad emakeelena eesti keelt (54,0%). Eesti keelt kõnelevate laenusaajate hulgas on häid kliente vähem (70,3%) kui vene keelt kõnelevate klientide hulgas (73,0%). Hii-ruut testi (H=3,3756 ja p-väärtus 0,06617) põhjal otsustati, et tunnused staatus ja keel ei ole küll statistiliselt omavahel sõltuvad, kuid valimi veidi teistmoodi võtmisel võib see tunnus oluliseks osutuda.

Laenusumma

Laenusumma keskmine on 324,3, miinimumväärtuseks 50 ja maksimumväärtuseks 2000 eurot. Jooniselt on näha, et laenude seas oli üksikuid suuri väärtusi, kuid enamik väärtusi jäi vahemikku 50 kuni 300 eurot. Kõige rohkem oli väärtusi vahemikus 200-300 eurot. Šansside logaritmide graafikult võib näha, et need, kelle laenusumma on suur või väike on suuremad šansid laenu korralikult tagasi maksta, kui nendel, kelle laenusumma on keskmine. Kuna enamik laenusummasid on kuni mõnisada eurot, siis võib kokkuvõttes oodata negatiivset seost.

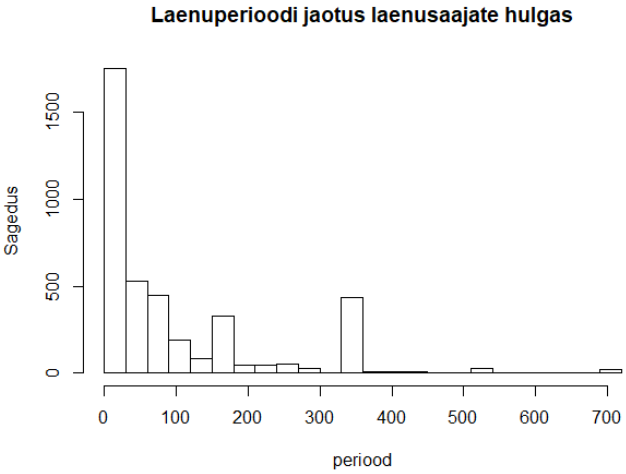


Joonis 2a. Šansi logaritmid erinevates summagruppides

Joonis 2b. Laenusumma jaotus laenusaajate hulgas

Periood

Keskmiseks laenuperioodiks on 112,7 päeva, miinimumväärtuseks 1 päev ja maksimumväärtuseks 720 päeva. Kõige populaarsemaks perioodiks oli 30 päeva (1 kuu). Samuti võeti palju laene 60, 90, 180 ja 360ks päevaks. Jooniselt 3 võib näha ka tunnuse periood täpsemat jaotust:



Joonis 3. Laenuperioodi jaotus laenusaajate hulgas

Vaadeldi ka kuidas on omavahel seotud staatus ja periood. Selleks jaotati inimesed perioodi alusel kahte gruppi: need kelle laenu periood oli rohkem kui 30 päeva ja teised, kellel see oli lühem või võrdne 30 päevaga.

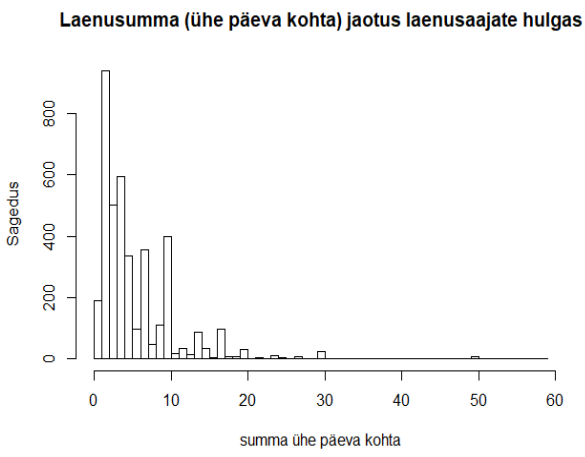
Tabel 4. Staatuse ja perioodi omavaheline jaotus

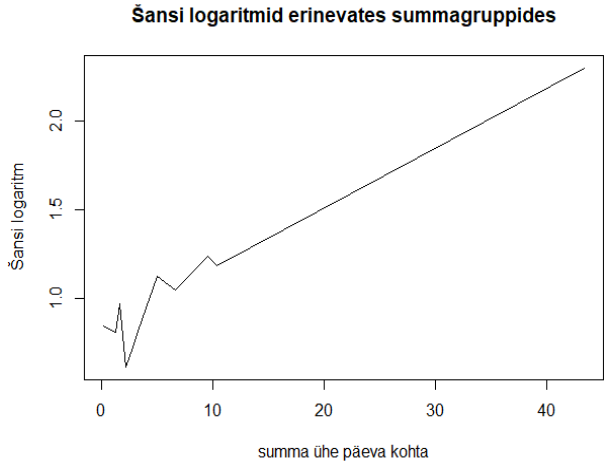
Staatus/periood	Periood<=30	Periood>30	Kokku
0	383 (21,9%)	756 (33,6%)	1139
1	1366 (78,1%)	1493 (66,4%)	2859
Kokku	1749	2249	3998

Tabelist 4 on näha, et paremad laenu tagasimaksjad on need, kelle laenu periood on lühem (78,1%). Nende seas, kelle laenu periood on rohkem kui 1 kuu on häid kliente 1493 (66,4%). Hii-ruut testi väärtuseks saadi 65,723 (p-väärtuse $5{,}189*10^{-16}$). Saab järeldada, et seos on oluline ja paremad kliendid on need, kelle laenu periood on lühem.

Laenusumma ühe päeva kohta

Logistilise mudeli koostamisel osutus oluliseks ka üks tunnus, mida algses andmestikus ei esinenud. Selleks oli laenusumma ja perioodi jagatis. Selliselt leitud laenusumma keskväärtuseks oli 6,0 eurot päevas, miinimumväärtuseks oli 0.24 eurot päevas ja maksimumväärtuseks 450 eurot päevas. Jooniselt 4a on näha, et enamik laenusummasid ühe päeva kohta jääb vahemikku 0-10 eurot. Šansi logaritmide graafikult (joonis 4b) on näha, et kliendid, kellel on suurem laenusumma päeva kohta maksavad paremini laene tagasi, kui need, kellel laenusumma päeva kohta on väiksem.



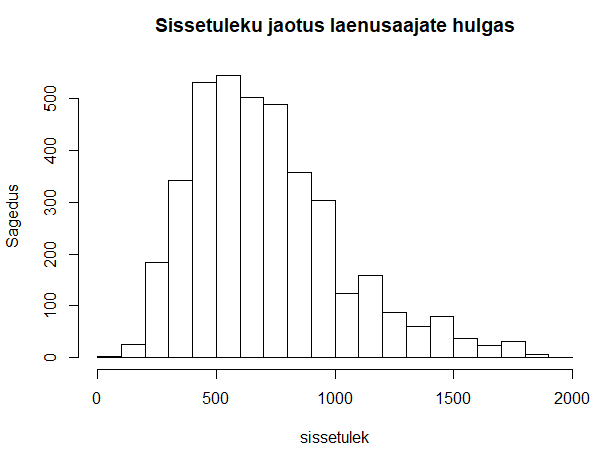


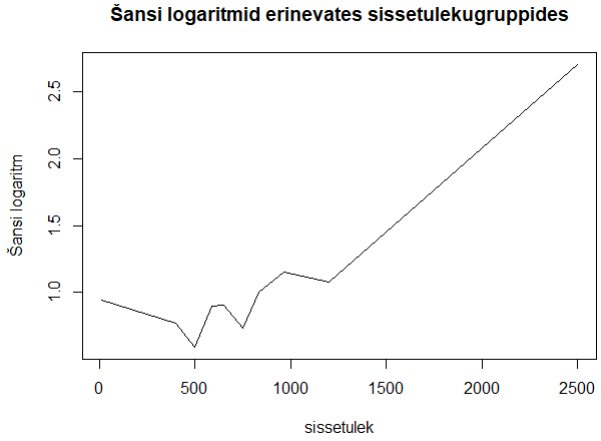
Joonis 4a. Laenusumma (ühe päeva kohta) jaotus laenusaajate hulgas

Joonis 4b. Šansi logaritmid erinevates summagruppides

Sissetulek

Laenusaajate keskmine sissetulek on 790,4 eurot, miinimumväärtuseks on 13,0 eurot ja maksimumväärtuseks 9500,0 eurot. Jooniselt 5a on näha, et kõige arvukamalt on inimestel sissetulekuid vahemikus 400-800 eurot. Šansi logaritmide graafikult (joonis 5b) võib näha, et üksteisele järjestikku olevate gruppide vahel esineb negatiivseid seos, kuid üldiselt on seos siiski nõrgalt positiivne.



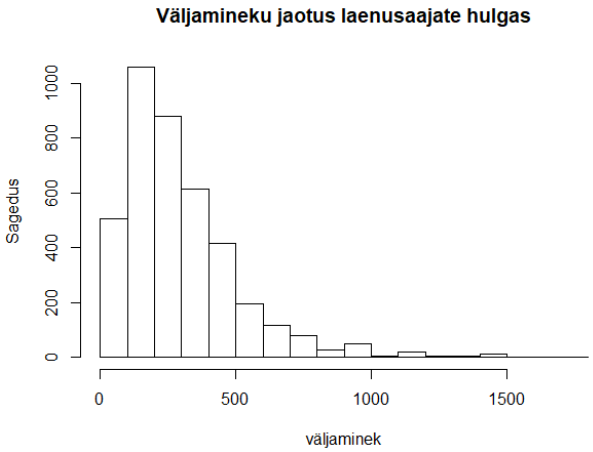


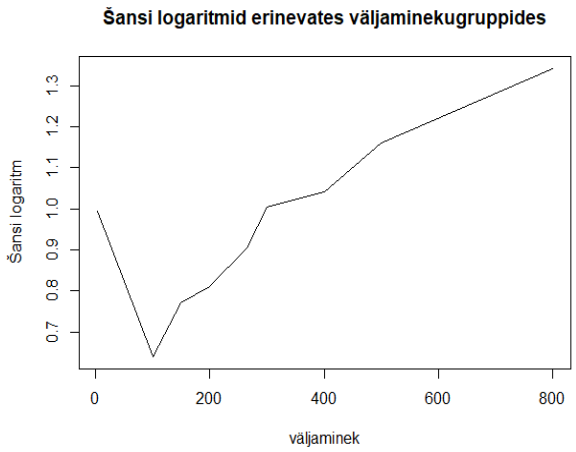
Joonis 5a. Sissetuleku jaotus laenusaajate hulgas

Joonis 5b. Šansi logaritmid erinevates sissetulekugruppides

Väljaminek

Laenusaajate keskmine väljaminek on 329,8 eurot, miinimumväärtuseks on 4,0 eurot ja maksimumväärtuseks 5000,0 eurot. Jooniselt 6a on näha, et enamik laenusaajate väljaminekud jäävad vahemikku 0-500 eurot ehk on pigem väikesed. Šansside logaritmide graafikult (joonis 6b) on näha, et kahe esimese grupi vahel, kellel on väljaminekud kõige väiksemad, on langev joon (seos on negatiivne). Kuid edasi on seos staatuse ja väljaminekute vahel laenusaajate seas juba positiivne ehk suurema väljaminekuga kliendid on laene paremini tagasi maksnud. Paistab, et väljaminek ja sissetulek on suhteliselt sarnaste näitajatega, kuid kuna sissetuleku lisamine mudelisse andis paremaid tulemusi, langes valik selle tunnuse kasuks.





Joonis 6a. Väljamineku jaotus laenusaajate hulgas

Joonis 6b. Šansi logaritmid erinevates väljaminekugruppides

Pereseis

Andmestikus oli kokku viis erineva pereseisuga kliente: abielus, vabaabielus, lesestunud, lahutatud ja vallalised. Laenusaajate pereseisu jaotus on ära toodud tabelis 5:

Tabel 5. Staatuse ja pereseisu omavaheline jaotus

Staatus/pereseis	Abielus	Lahutatud	Lesk	Vabaabielus	Vallaline	Kokku
0	273 (23,5%)	90 (21,5%)	24 (19,0%)	337 (22,0%)	415 (33,5%)	1139
1	889 (76,5%)	329 (78,5%)	102 (81,0%)	715 (68,0%)	824 (66,5%)	2859
Kokku	1162	419	126	1052	1239	3998

On näha, et suhteliselt võrdselt on valimis abielus, vabaabielus ja vallalisi inimesi. Vähem on lahutatud ja leski. Halvema maksekäitumisega on vabaabielus ja vallalised laenusaajad, vastavalt 68% ja 66,5% on hea staatusega kliendid. Kõige parema maksekäitumisega on lesestunud laenusaajad, kuid kuna neid on valimis vähe, siis suuremat tähtsust omab see, et kolme suuremat pereseisu gruppi võrreldes näeme, et nendest on abielus kliendid kõige paremini laenu tagasi maksnud. Hii-ruut test andis statistiku väärtuseks 51,577 (p-väärtusega $1{,}691*10^{-10}$). Seega võib öelda, et tehtud järeldused on ka statistiliselt olulised ja staatus ning pereseis on omavahel sõltuvad.

Haridus

Andmestikus oli haridusel kuus taset: kõrgharidus, keskharidus, kutseharidus, põhialgharidus, algharidus ja need, kellel see puudub. Laenusaajate haridustaseme jaotus on ära toodud tabelis 6:

Tabel 6. Staatuse ja haridustaseme omavaheline jaotus

Staatus/ haridustase	Kõrg-haridus	Kesk-haridus	Kutse-haridus	Põhi-haridus	Alg-haridus	Ei ole	Kokku
0	131 (17,9%)	462 (29,8%)	322 (26,7%)	197 (45,3%)	11 (35,5%)	16 (22,7%)	1139
1	599 (82,1%)	1086 (70,2%)	883 (73,3%)	238 (54,7%)	20 (64,5%)	33 (67,3%)	2859
Kokku	730	1548	1205	435	31	49	3998

Kõige rohkem on valimisse sattunud laenusaajate hulgas kõrg- ja kutseharidusega kliente. Kokku moodustavad nad ligi 70% klientidest. Kui mitte arvestada algharidusega inimesi ja neid, kes on märkinud hariduseks „ei ole“ (neid on valimis väga vähe), siis võib öelda, et mida kõrgem on haridus, seda suurema tõenäosusega on tegemist staatuselt hea kliendiga. Hii-ruut testiga kontrolliti, kas see oletatav seos on ka statistiliselt oluline. Väärtuseks saadi 104,49 (p-väärtus < $2{,}2*10^{-16}$), seega on haridus ja staatus omavahel sõltuvad.

Töökogemus

Laenusaajad olid nelja erinevat tüüpi töökogemusega: töötu, katseajal, kuni aasta ja rohkem kui aasta. Töökogemuse jaotus on ära toodud tabelis 7:

Tabel 7. Töökogemuse ja staatuse omavaheline jaotus

Staatus/ töökogemus	Katseaeg	Kuni aasta	Rohkem kui aasta	Töötu	Kokku
0	34 (33,7%)	277 (42,4%)	780 (25,5%)	48 (26,8%)	1139
1	67 (66,3%)	377 (57,6%)	2284 (74,5%)	131 (73,2%)	2859
Kokku	101	654	3064	179	3998

On näha, et ligikaudu kolmveerandil klientidel on töökogemust rohkem kui aasta. Kõige suurema töökogemusega kliendid maksavad ka laene kõige paremini tagasi (häid kliente on 74,5%). Umbes sama hea maksekäitumisega on ka töötud laenusaajad (73,2%). Kui võrrelda kahte suuremat gruppi töökogemuse järgi, siis võib näha, et mida suurem on töökogemus, seda parema maksekäitumisega on kliendid. Vaadeldi, kas see seos on ka statistiliselt oluline. Hii-ruut testi väärtuseks saadi 77,118 (p-väärtusega $< 2{,}2*10^{-16}$). Seega on staatus ja töökogemus omavahel statistiliselt sõltuvad.

Laste arv

Andmestikus oli laste arvu varieeruvus laenusaajatel 0-10. Tabelis 8 on välja toodud laenusaajad, kellel ei olnud lapsi ja kelle laste arv oli 1, 2, 3 või rohkem kui 3:

Tabel 8. Staatuse ja laste arvu omavaheline jaotus

Staatus/lapsi	0	1	2	3	>3	Kokku
0	682 (27,4%)	262 (29,3%)	139 (30,3%)	45 (33,8%)	11 (33,3%)	1139
1	1805 (72,6%)	633 (70,7%)	311 (69,7%)	88 (66,2%)	22 (66,7%)	2859
Kokku	2487	895	450	133	33	3998

Kõige rohkem on andmestikus laenusaajaid, kellel ei olnud lapsi, neid on 62,2% kogu valimi mahust. Hea maksekäitumisega klientide protsente vaadeldes näib, et mida väiksem on laenusaaja laste arv, seda paremini maksab klient laenu tagasi. Seda seaduspära hii-ruut testiga kontrollides saadi tulemuseks 5,1765 (p-väärtusega 0,2697). Järelikult ei saa tõestada, et staatus ja laste arv on omavahel sõltuvad ja näiv seaduspära osutus statistiliselt mitteoluliseks.

Kinnisvarade arv

Valimisse sattunud klientidel varieerus kinnisvarade arv nullist kinnisvarast kaheksa kinnisvarani. Tabelis 9 on näha klientide kinnisvarade arvu jaotus:

Tabel 9. Staatuse ja kinnisvarade arvu omavaheline jaotus

Staatus/kinnisvarasid	0	1	2	>2	Kokku
0	816 (37,4%)	267 (19,3%)	43 (13,4%)	13 (12,0%)	1139
1	1368 (62,6%)	1117 (80,7%)	279 (86,6%)	95 (88,0%)	2859
Kokku	2184	1384	322	108	3998

Natuke rohkem kui pooled laenusaajatest ei oma ühtegi kinnisvara. Need laenusaajad, kelle omanduses ei ole ühtegi kinnisvara, on halvemad laenu tagasimaksjad, kui need, kes omavad kinnisvara: heade klientide protsendid kinnisvara mitte omavatel klientidel on 62,7% ja kinnisvara omavatel klientidel üle 80%. Heade klientide protsente vaadeldes paistab välja seaduspära, et mida suurem on kliendi kinnisvarade arv, seda paremini maksab ta laenu tagasi. Hii-ruut testiga kontrollides saadi tulemuseks 192,43 (p-väärtusega $< 2{,}2*10^{-16}$). Seega on staatus ja kinnisvarade arv omavahel sõltuvad.

Maksehäired

Andmestikus olid omavahel eristatud lõpetatud ja aktiivsed maksehäired ja vaadeldi ka seda, kui palju oli kliendil maksehäireid kokku. Tabelis 10 on ära toodud maksehäirete jaotus klientide hulgas:

Tabel 10. Staatuse ja maksehäirete omavaheline jaotus

Staatus/ maksehäireid	Aktiivsed		Suletud		Aktiivsed + suletud	
	Ei ole	On	Ei ole	On	Ei ole	On
0	966 (27,1%)	173 (39,8%)	581 (23,5%)	558 (37,5%)	528 (22,6%)	611 (36,8%)
1	2597 (72,9%)	262 (60,2%)	1889 (76,5%)	970 (63,5%)	1811 (77,4%)	1048 (63,2%)
Kokku	3563	435	2470	1528	2339	1659

On näha, et kliendid, kellel ei ole maksehäireid, maksavad laenu tagasi paremini, kui kliendid, kellel neid leidub. Kõige paremini on seda vahet märgata, kui vaadelda maksehäireid kokku ehk klientide hulgas, kellel ei ole aktiivseid ega suletuid maksehäireid, on häid kliente 77,4% ja nende seas, kellel leidub aktiivseid või lõpetatud maksehäireid, on häid kliente 63,2%. Iga grupi seas vaadeldi eraldi hii-ruut testiga, kas sõltuvus staatuse ja maksehäirete olemasolu vahel on olemas. Saadi, et aktiivsete maksehäirete olemasolu, lõpetatud maksehäirete olemasolu ning ühe või teise maksehäire olemasolu ja staatuse vahel on seos olemas (teststatistiku väärtus oli suurem (vastavalt 29,87, 77,63 ja 97,12) kui kriitiline piir ühe vabadusastme korral (3,84)).

3.3 Logistilise regressiooni mudeli loomine

Logistilise regressiooni mudeli loomiseks kasutati statistikaprogrammi R funktsiooni „glm“ (logistiline regressioon on üldine lineaarne mudel) ja funktsiooni argumendina lisati, et prognoositakse 0/1 tunnust (family=„binomial“).

3.3.1 Logistiline regressioon lähteandmestikul

Kõigepealt loodi mudel algse valimi põhjal kasutades ettepoole sammregressiooni, kus alustati ainult vabaliikmega mudelist ja lisatakse järjest argumente, võrreldes nende p-väärtusi mudelisse lisamisel. Selliselt käitudes saadi krediidiriski hindamiseks järgmine mudel:

*Logit(π) = 2,5587 + 1,0287 * ln(kinnisvarade arv + 1) + 0,0244 * vanus + 0,3462 * (haridus=kõrgharidus) + 0,3383 * (maakond=Harjumaa) – 0,6835 * ln(laenusumma) - 0,4901 * (maksehäireid kokku=on olemas) + 0,0456 * (laenusumma päeva kohta) + 0,0024 * laenuperiood +0,0006 * väljaminek - 0,2218 * (sugu=mees) - 0,1581 * (aktiivseid maksehäireid) + 0,1117 * töökogemus*,

kus π tähistab laenusaaja tõenäosust olla hea staatusega klient.

Esialgset tunnust haridus modifitseeriti nii, et vaadeldi, kas inimesel on kõrgharidus või mitte. Maakonna puhul vaadeldi, kas inimene on Harju maakonnast või mitte. Kinnisvarade arvu vaadeldi nii nagu ka selle tunnuse esmasel kirjeldamisel ehk laenusaajad jagati gruppidesse: ei oma kinnisvara, omab 1 kinnisvara, 2 kinnisvara, rohkem kui 2 kinnisvara. Logaritmi võtmiseks pidi tunnuse väärtustele liitma 1. Maksehäirete kogusumma asemel vaadeldi, kas laenusaajal on esinenud aktiivseid või lõpetatud maksehäireid või mitte. Töökogemust kodeeriti järgnevalt: töötutele inimestele anti töökogemuse väärtuseks 0, katseajal olevatele inimestele 1, töökogemusega kuni aasta 2 ja töökogemus rohkem kui aasta sai väärtuseks 3. Ülejäänud tunnuste puhul kasutati nende algset või logaritmitud kuju. Logaritmimist kasutati seetõttu, et kui tunnusel esineb mõõdukas saba paremale, siis soovitatakse võtta tunnuse väärtusest naturaallogaritm, mis vähendab tunnuse hajuvust ja muudab jaotust sümmeetrilisemaks. [4]

Sooviti vaadelda, kui täpselt suudab loodud mudel klientide staatust. Selleks valiti katsetamise teel otsustuspiiriks 0,488. Prognoosid, mis ületasid otsustuspiiri, said väärtuseks 1 ja

prognoosid, mis jäid alla otsustuspiiri, said väärtuseks 0. Tabelist 11 on näha, kui täpselt suutis mudel ennustada halbade ja heade klientide staatust.

Tabel 11. Tasakaalustatumata andmestiku põhjal loodud mudeli täpsus

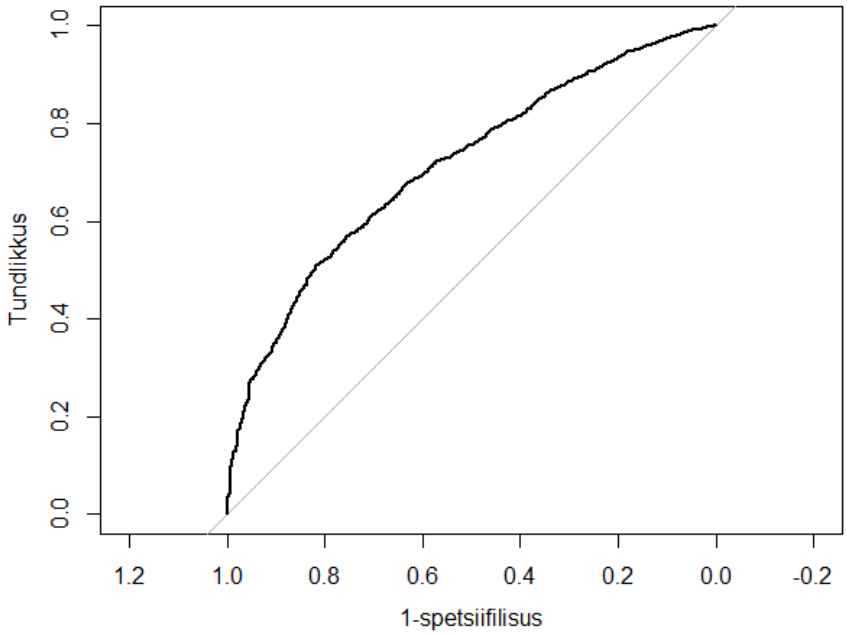
Tegelik/prognoos	0	1	Kokku
0	208	931	1139
1	155	2704	2859
Kokku	363	3635	3998

On näha, et mudel prognoosis väga hästi heade klientide staatust. Täpselt prognoositud hea staatusega klientide protsent kõikidest tegelikult hea staatusega klientidest on 94,6%. Seevastu ei saanud mudel hakkama halbade klientide prognoosimisega. Kõikidest halva staatusega klientidest suudeti täpselt prognoosida 208, mis on 18,3% tegelikust halva staatusega klientide arvust. Paistab, et prognoos on heade klientide poole kallutatud, kuna prognoosi järgi oli häid kliente 90,1%, kuid tegelikus andmestikus oli neid 71,5%. Kõikide õigesti prognoositud klientide arv oli 2912, mis moodustab 72,8% kogu vaatluste arvust. Selliste näitajatega mudeli kasutusulevõtmine mingi finantsasutuse poolt võib põhjustada suuri finantskulusid, sest kõrge riskiastmega kliendid omavad väiksemat tõenäosust laenu korralikult tagasi maksta ja antud mudel ennustas halbade klientide staatust kehvasti.

Tabel 12. Tasakaalustamata andmete põhjal loodud mudeli headuse näitajad

Statistik	ROC-kõvera alune pindala	Akaike informatsioonikordaja	Ruutkeskmine viga
Väärtus	0,7132	4331,2	0,1812

Tabelist 12 on näha, et loodud mudeli Akaike informatsioonikordajaks oli 4331,2 ja ruutkeskmiseks veaks 0,1812. ROC-kõvera aluseks pindalaks oli 0,7132. Joonisel 7 on kuvatud mudeli ROC-kõver.



Joonis 7. Tasakaalustatumata andmestiku põhjal loodud mudeli ROC-kõver

3.3.2 Logistiline regressioon tasakaalustatud andmetel

Kuna loodud mudel prognoosis halvasti halbade klientide staatust, kasutati ka andmete tasakaalustamist.

Andmete tasakaalustamisel võeti uude valimisse kõik halva staatusega kliendid ja nendele lisati võrdne arv hea staatusega kliente. Hea staatusega laenusaajate valimisse lisamisel kasutati lihtsat juhuslikku valikut. Et kontrollida, kas andmete tasakaalustamine on korrektne, lisati esialgsesse andmestiku enne tasakaalustamist tunnus ennustus1, mis näitas seda, kuidas ennustas algandmestiku poolt loodud mudel uuritavat tunnust staatus. Järgmiseks loodi samade argumentidega mudel uue valimi põhjal ja lisati tunnus ennustus2, mis näitas uue valimi põhjal loodud mudeli ennustusi. Seejärel leiti Spearmani korrelatsioonikordaja ennustus1 ja ennustus2 vahel, et kontrollida, kas mudeli põhjal saadud ennustuste (ennustus2) järjestus on jäänud samaks, kui esialgse mudeli korral (ennustus1). Oodatavaks korrelatsioonikordaja väärtuseks oli ligikaudu 1, kordaja ennustus1 ja ennustus2 vahel andis tulemuseks 0.968. Tasakaalustatud andmestikus oli 2278 vaatlust.

Lõplik mudel loodi kasutades ettepoole regressiooni. Võrdluseks prooviti ka mudeli loomist kasutades tahapoole sammregressiooni, kus alustati täismudelist ja hakati järjest p-väärtuse põhjal tunnuseid välja jätma kuni saadi mudel, kus kõik tunnused on olulised. Ette- ja tahapoole

sammregressiooniga koostatud mudelite headuse näitajad (ROC-kõvera alune pindala, Akaike informatsioonikriteerium (AIC) ja ruutkeskmine viga) on ära toodud tabelis 13.

Tabel 13. Erinevate sammregressiooni tüüpidega loodud mudelite võrdlus

Sammregressiooni tüüp/headuse näitaja	ROC-kõvera alune pindala	Akaike informatsioonikordaja	Ruutkeskmine viga
Ettepoole	0,7208	2824,9	0,2127
Tahapoole	0.7038	2890,2	0,2185

On märgata, et ettepoole sammregressiooniga loodud mudel edestas headuse näitajate poolest tahapoole sammregressiooniga loodud mudelit. ROC- kõvera alune pindala on suurem, Akaike informatsioonikordaja väiksem ja ruutkeskmine viga samuti väiksem.

Ka tasakaalustatud andmestiku põhjal loodud mudeli puhul vaadeldi mudeli prognoosi täpsust. Leiti otsustuspiir, mille järgi jagati kliendid prognoosi järgi halbadeks ja headeks klientideks. Katsetamise teel valiti otsustuspiiriks 0,475.

Tabel 14. Tasakaalustatud andmestiku põhjal loodud mudeli täpsus

Tegelik/prognoos	0	1	Kokku
0	736	403	1139
1	362	777	1139
Kokku	1098	1180	2278

Tabelist 14 on näha, et tasakaalustatud andmestiku põhjal loodud logistilise regressioonimudeli korral on kogu õigesti prognoositud vaatluste protsent kogu andmestikus olevatest vaatlustest küll väiksem kui tasakaalustamata andmestiku puhul (tasakaalustatud andmestiku puhul 66,4% ja tasakaalustamata andmestiku puhul 72,8%), kuid tasakaalustatud andmestiku põhjal loodud mudel prognoosib märgatavalt paremini halbade klientide staatust (õigesti prognoositakse 66,4% juhtudel, sama näitaja tasakaalustamata mudeli puhul oli 18,3%). Samuti on märgata, et tasakaalustatud mudeli põhjal ei ole mudel kallutatud: kui andmestikus oli hea ja halva staatusega kliente võrdselt, siis mudeli prognoosi järgi oli häid kliente 51,8% ja halbu kliente 48,2%.

Võrreldi ka tasakaalustatud ja tasakaalustamata andmestiku ruutkeskmist viga. Pelgalt statistiku väärtusi võrreldes selgub, et tasakaalustamata andmestiku põhjal loodud mudel on täpsem ehk tema ruutkeskmine viga on väiksem. Uuriti ka ruutkeskmist viga eraldi halbade ja heade klientide seas. Järeldused olid sarnased sagedustabeli põhjal tehtutele: tasakaalustatud mudeli põhjal loodud logistilise regressiooni mudel prognoosis paremini halbade klientide staatust (tasakaalustatud mudeli puhul oli ruutkeskmine viga 0,2084 ja tasakaalustamata mudeli puhul 0,4243) ning halvemini heade klientide staatust (ruutkeskmised vead vastavalt 0,2167 ja 0,0843).

Lisaks sellele, et tasakaalustatud andmestiku põhjal loodud regressioonmudel prognoosis paremini halbade klientide staatust, oli ka selle mudel ROC-kõvera alune pindala suurem kui tasakaalustamata mudeli puhul. Tasakaalustatud andmestiku puhul loodud regressioonmudel prognoosis paremini halbade klientide staatust kui tasakaalustamata andmete põhjal loodud mudel. Seega tasakaalustatud andmete põhjal loodud mudeli kasutamine ei põhjusta nii suuri finantskulusid kui tasakaalustamata andmete põhjal loodud regressioonmudel, sest kõrge riskiastmega kliente suudetakse paremini klassifitseerida. Seega võib eelistada tasakaalustatud andmete põhjal loodud logistilise regressiooni mudelit ja öelda, et andmete tasakaalustamine oli õigustatud. Paremaks osutunud mudelit uuriti ka lähemalt.

Tabel 15. Mudelis oluliseks osutunud tunnuste kordajad

Tunnus/kordaja	Hinnang	p-väärus
Vabaliige	1,6877	0,0005
Ln(kinnisvarade arv +1)	1,0828	$<2*10^{-16}$
Vanus	0,0269	$2,66*10^{-11}$
Haridus=kõrgharidus	0,4253	0,0011
Maakond=Harjumaa	0,3154	0,0008
Ln(laenusumma)	-0,6827	$9,56*10^{-12}$
Maksehäired kokku=on olemaas	-0,5132	$1,09*10^{-7}$
Laenusumma päeva kohta	0,0406	0,0008
Laenuperiood	0,0027	$5,75*10^{-6}$
Sissetulek	0,0003	0,0197
Sugu=mees	-0,2401	0,0141
Aktiivseid maksehäireid	-0,2189	0,0009
Emakeel=vene	0,2292	0,0140

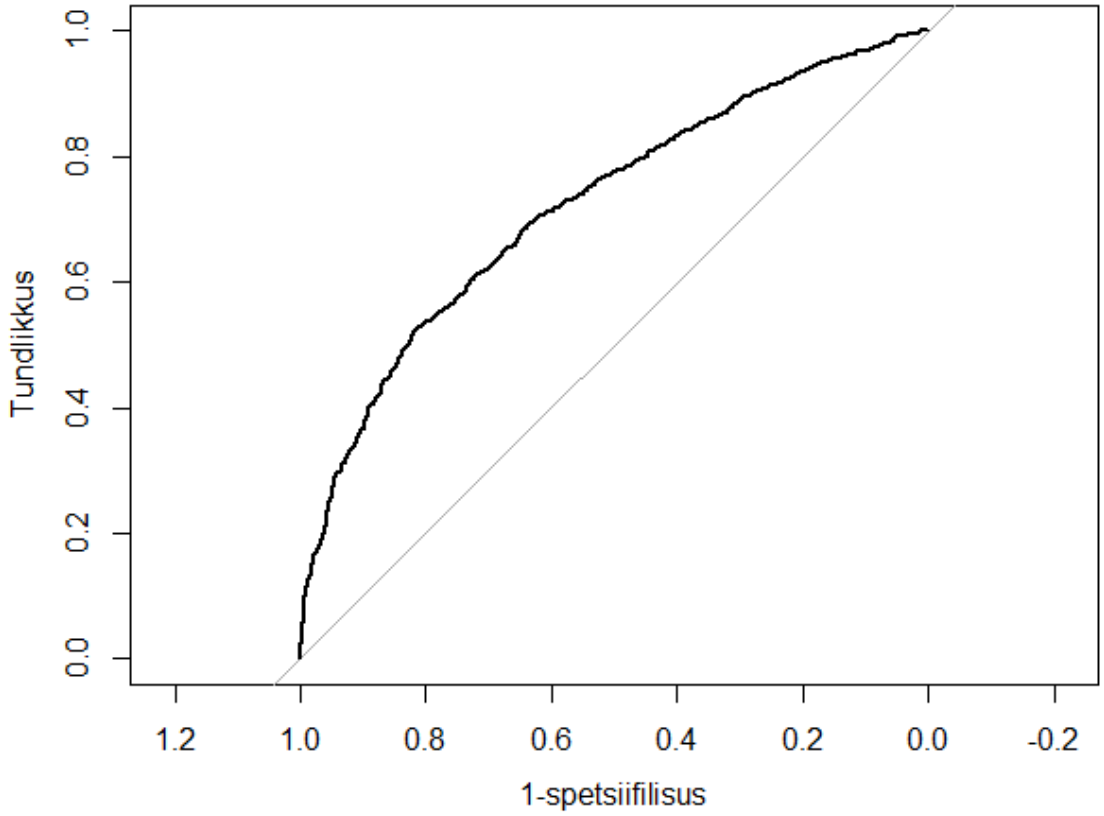
Tabelist 15 on näha, et mudelis osutusid oluliseks 12 tunnust. Võrreldes tasakaalustamata mudeliga on mudelis väljaminekute asemel sissetulek ja töökogemuse asemel osutus mudelis oluliseks laenusaaja räägitav emakeel.

Logistilisest regressioonist saadi laenusaajate krediidiriski hindamiseks järgmine mudel:

*Logit(π) = 1,6877 + 1,0828 * ln(kinnisvarade arv + 1) + 0,0269 * vanus + 0,4253 * (haridus=kõrgharidus) + 0,3154 * (maakond=Harjumaa) – 0,6827 * ln(laenusumma) - 0,5132 * (maksehäireid kokku=on olemas) + 0,0406 * (laenusumma päeva kohta) + 0,0027 * laenuperiood +0,003 * sissetulek - 0,2401 * (sugu=mees) - 0,2189 * (aktiivseid maksehäireid) + 0,2292 * (emakeel=vene),*

kus π tähistab laenusaaja tõenäosust olla hea staatusega klient.

Joonisel 8 on kuvatud mudeli ROC-kõver. Mudeli ROC-kõvera aluseks pindalaks on 0,7208. Järelikult on tegu aktsepteeritava mudeliga. Hosmer - Lemeshow' test andis teststatistiku väärtuseks 11,064 ja olulise tõenäosuseks 0,1964, seega ei saa lükata ümber nullhüpoteesi, et mudel sobib andmetega.



Joonis 8. Laenusaajate mudeli ROC-kõver tasakaalustatud andmete korral

3.4 Mudeli interpretatsioon

Šansside võrdlused on tehtud eeldusel, et teiste tunnuste väärtused ei muutu ehk muud tingimused peale käsitletava tunnuse muutumise jäävad samaks. Tunnuseid käsitletakse mudelis esinemise järjekorras.

Kui võrrelda kahte laenusaajat, siis klient, kellel on ühe võrra rohkem kinnisvarasid, omab $2{,}953^{(\ln(a+1)-\ln(a))}$ korda ($e^{1{,}0828*(\ln(a+1)-\ln(a))}$)=$2{,}953^{(\ln(a+1)-\ln(a))}$) suuremaid šansse osutuda heaks laenu tagasimaksjaks. (a>1)

Laenusaajal, kellel on vanus ühe võrra suurem kui teisel laenusaajal, omab 2,7% ($e^{0{,}0269}\approx1{,}027$) suuremaid šansse laenu korralikult tagasi maksta.

Kui võrrelda kaht inimest, kellest ühel on kõrgharidus ja teisel mitte, siis kõrgharidusega laenusaajatel on 53% ($e^{0.4253}\approx1{,}530$) paremad šansid osutuda hea maksekäitumisega laenusaajaks.

Kliendil, kes elab Harju maakonnas, on 37,1% ($e^{0{,}3154}\approx1{,}371$) suuremad šansid laenu korralikult tagasi maksta kui kliendil, kes ei ela Harju maakonnas.

Mudelist paistab ka see, et kui võrrelda kahte klienti, kellest ühe laenusumma on teisest ühe euro võrra suurem, siis tema šansid laenu korralikult tagasi maksta on $1/(0{,}505^{(\ln(a+1)-\ln(a))})$ (($e^{-0{,}6827}$)$^{(\ln(a+1)-\ln(a))}\approx0{,}505^{(\ln(a+1)-\ln(a))}$) korda väiksemad. (a – kliendi laenusumma (a>=50))

Laenusaajal, kellel leidub kas suletud või aktiivseid maksehäireid, on 66,9 ($e^{-0{,}5132}\approx0{,}599$; $1/0{,}599 \approx 1{,}669$) väiksemad šansid osutuda hea maksekäitumisega laenusaajaks kui laenusaajal kellel ei leidu suletud ega aktiivseid maksehäireid.

Klient, kelle laenusumma päeva kohta on teisest kliendist ühe euro võrra suurem, omab 4,1% ($e^{0{,}0406}\approx1{,}041$) suuremaid šansse laenu korralikult tagasi maksta.

Laenusaajal, kelle laenuperiood on teisest ühe päeva võrra pikem, on 0,3% ($e^{0{,}0027}\approx1{,}003$) suuremad šansid osutuda heaks laenu tagasimaksjaks.

Mudelist ilmneb, et kliendil, kelle sissetulek on teisest ühe euro võrra suurem, on 0,3% ($e^{0{,}003}\approx1{,}003$) suuremad šansid olla hea maksekäitumisega laenusaaja.

Laenusaaja, kes on meessoost, omab 27,1% ($e^{-0{,}2401}\approx0{,}787$; $1/0{,}787 \approx 1{,}271$) väiksemaid šansse osutuda korralikuks laenusaajaks kui naissoost laenusaaja.

Kliendil, kellel on teisest ühe võrra rohkem aktiivseid maksehäireid, on 24,4% ($e^{-0,2189} \approx 0,803$; $1/0,803 \approx 1,244$) väiksemad šansid olla hea maksekäitumisega klient.

Mudelist paistab, et laenusaaja, kelle emakeeleks on vene keel, omab 25,8 % ($e^{0,2292} \approx 1,258$) suuremaid šansse osutuda korralikuks laenu tagasimaksjaks kui laenusaaja, kelle emakeeleks on eesti keel.

Kokkuvõtteks võib öelda, et kõige suuremad šansid osutuda hea maksekäitumisega kliendiks on laenusaajatel, kellel on rohkem kui kaks kinnisvara, kelle vanus on keskmisest kõrgem, kes omavad kõrgharidust, elavad Harju maakonnas, kelle laenusumma on keskmisest väiksem, kes ei oma suletud ega aktiivseid maksehäireid, kelle laenusumma päeva kohta on keskmisest suurem, kelle laenuperiood on keskmisest suurem, kes omab keskmisest suuremat sissetulekut, on naine ja räägib emakeelena vene keelt.

Kokkuvõte

Antud bakalaureusetöö eesmärgiks on logistilise regressiooni abil mudeli loomine laenusaajate andmete põhjal, et prognoosida tõenäosusi, et klient osutub heaks kliendiks. Teiseks eesmärgiks oli vaadelda, kuidas mõjutab prognoose andmete tasakaalustamine. Töös anti ülevaade krediidiriski olemusest ja kasutatud metoodikast. Samuti tehti esmaanalüüs tunnustele, et aimu saada, mis tunnused ja mis määral prognoositavaid tõenäosusi võiks mõjutada. Logistilise regressiooni abil loodi mudel krediidiriski hindamiseks nii tasakaalustamata kui ka tasakaalustatud andmete põhjal.

Laenusaajate andmete põhjal õnnestus välja töötada aktsepteeritavate kvaliteedinäitajatega mudel. Lõplikus mudelis osutusid oluliseks 12 tunnust. Mudeli kordajaid interpreteerides õnnestus välja selgitada, et kõige madalama krediidiriskiga kliendiks on keskmisest vanem vene keelt kõnelev naine, kes omab kõrgharidust, elab Harju maakonnas, omab keskmisest suuremaid väljaminekuid ja tal ei ole olnud ei aktiivseid ega suletud maksehäireid. Samuti on tema laenuperiood keskmisest pikem ja summa ühe päeva kohta suurem.

Logistiline regressioon on parameetriline meetod, täpsemalt on tegu üldistatud lineaarse mudeli erijuhuga, kus seosefunktsioon on $logit$-funktsioon. Algselt rakendati töös logistilise regressiooni mudelit kogu andmestiku peal. Selgus, et selle mudeli headuse näitajad ei olnud piisavalt head ja mudel prognoosib suhteliselt halvasti kõrge riskiastmega laenusaajaid. Andmed tasakaalustati ja balansseeritud andmestiku põhjal alustati uue mudeli loomist. Kasutati nii ettepoole kui tahapoole regressiooni. Selgus, et ettepoole regressiooniga loodud mudel oli oma headuse näitajate poolest parem.

Tasakaalustatud ja tasakaalustamata mudelite võrdluseks kasutati kahemõõtmelist sagedustabelit, kus vaadeldi kui täpsed on kummagi mudeli prognoosid võrreldes tegelike väärtustega. Selgus, et balansseerimine suurendas märgatavalt halbade klientide prognoosimise täpsust ja vähendas heade klientide prognoosimise täpsust. Balansseerimine vähendas ka mudeli kallutatust. Tasakaalustamise kasuks rääkis ka see, et selle mudeli ROC-kõvera alune pindala oli suurem kui tasakaalustamata mudelil. Balansseeritud andmete põhjal loodud mudel osutus paremaks ja selle kordajaid ka interpreteeriti.

Kasutatud kirjandus

[1] Hand, D.J., Henley, W.E. (1997). Statistical Classification Methods in Consumer Credit Scoring: a Review. Journal of the Royal Statistical Society: Series A (Statistics in Society), Vol. 160, No. 3, 523-541

[2] Siddiqi, N. (2005). Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring. New Jersey: John Wiley & Sons, Inc.

[3] Kočenda, E., Vojtek, M. (2009). Default Predictors and Credit Scoring Models for Retail Banking. CESifo Working Paper, No. 2862.

[4] Käärik, E. (2014). Andmeanalüüs II. Loengukonspekt. Tartu: Tartu Ülikool, matemaatika ja statistika instituut.

Saadaval: http://dspace.ut.ee/bitstream/handle/10062/35401/AndmeanaluusII.pdf

[5] Hosmer, D.W., Lemeshow, S., Sturdivant, R.X. (2013). Applied Logistic Regression. New

Jersey: John Wiley & Sons, Inc.

[6] Agresti, A. (2002). Categorical Data Analysis. New Jersey: John Wiley & Sons, Inc.

[7] Salas-Eljatiba, C., Fuentes-Ramireza, A., Gregoireb, T. G., Altamiranoc, A., Yaitula, V. (2017). A study on the effects of unbalanced data when fitting logistic regression models in ecology. Ecological Indicators. 85. 10.1016/j.ecolind.2017.10.030.

[8] Crone, S., Finlay, S. (2012). Instance sampling in credit scoring: An empirical study of sample size and balancing. International Journal of Forecasting. 28. 224–238. 10.1016/j.ijforecast.2011.07.006.

[9] Kleinbaum, D.G., Kupper, L.L., Muller, K.E., Nizam, A. (1998). Applied Regression Analysis and Other Multivariable Methods. Duxbury Press.

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Indrek Polding,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose „Krediidiriski hindamine logistilise regressiooni mudeli abil", mille juhendaja on Kalev Pärna,

    1.1. reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace-is lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;

    1.2. üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace'i kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.

2. olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile.

3. kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Tartus, 08.05.2018